Tokyo Tech at TRECVID 2020: Relation Modeling for Video Action Detection

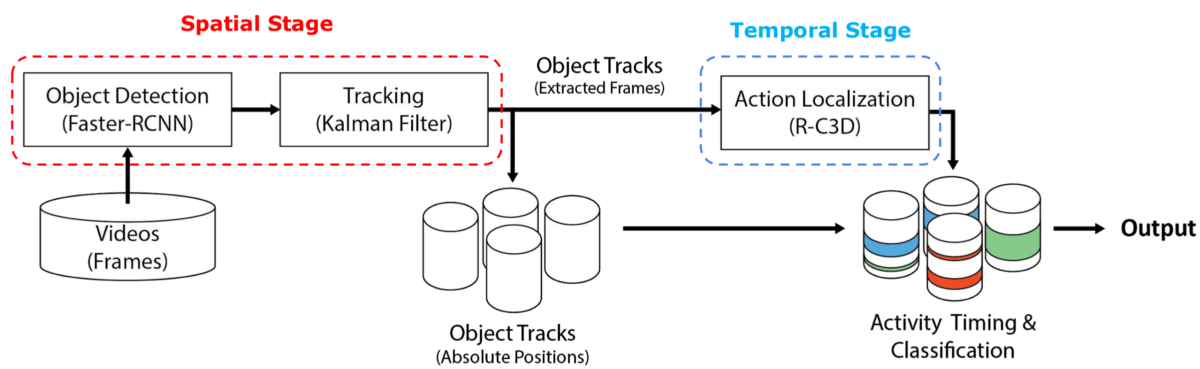
TokyoTech_AIST

Ronaldo Prata Amorim,

Nakamasa Inoue, Koichi Shinoda

Introduction

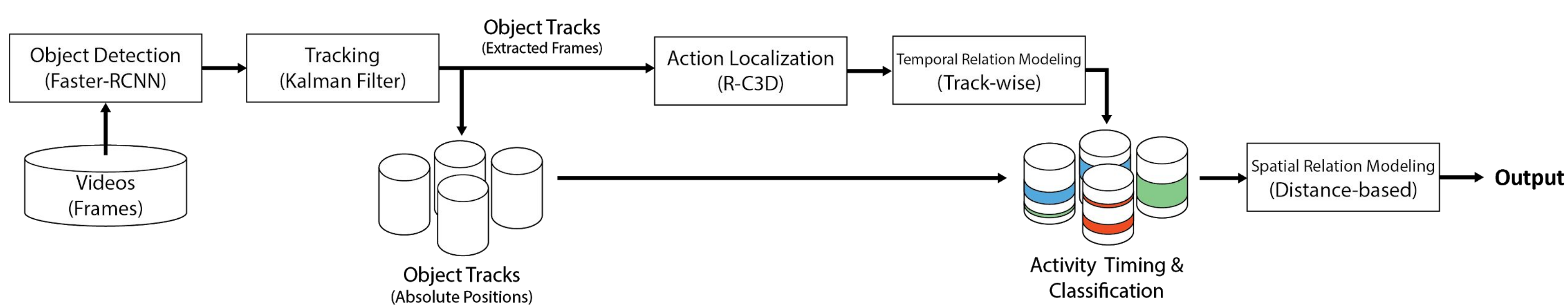
- Hard to process untrimmed, arbitrarily long videos in their entirety
- Fair recent success of two-stage spatial-temporal separating frameworks
- Separation leads to loss of information in individual stages
  - Spatial stage doesn't discern when objects are involved in actions
  - Temporal stage for isolated objects loses contextual information



Introduction

System Overview

- Two stage based framework
  - Spatial stage through frame-wise object detection
  - Temporal stage through object-wise action localization
- Relation modeling heuristics post-processing
  - Modeling temporal sequences of proposals of the same object
  - Modeling spatial distance between proposals of different objects



Tokyo Tech

System Framework

Object Detection and Tracking

- Frame-wise object detection (Faster-RCNN)
  - Person and vehicle objects (actors)
  - Spatial localization and classification every 5 frames
- Kalman Filter based object tracking
  - Object tracks for each detected actor

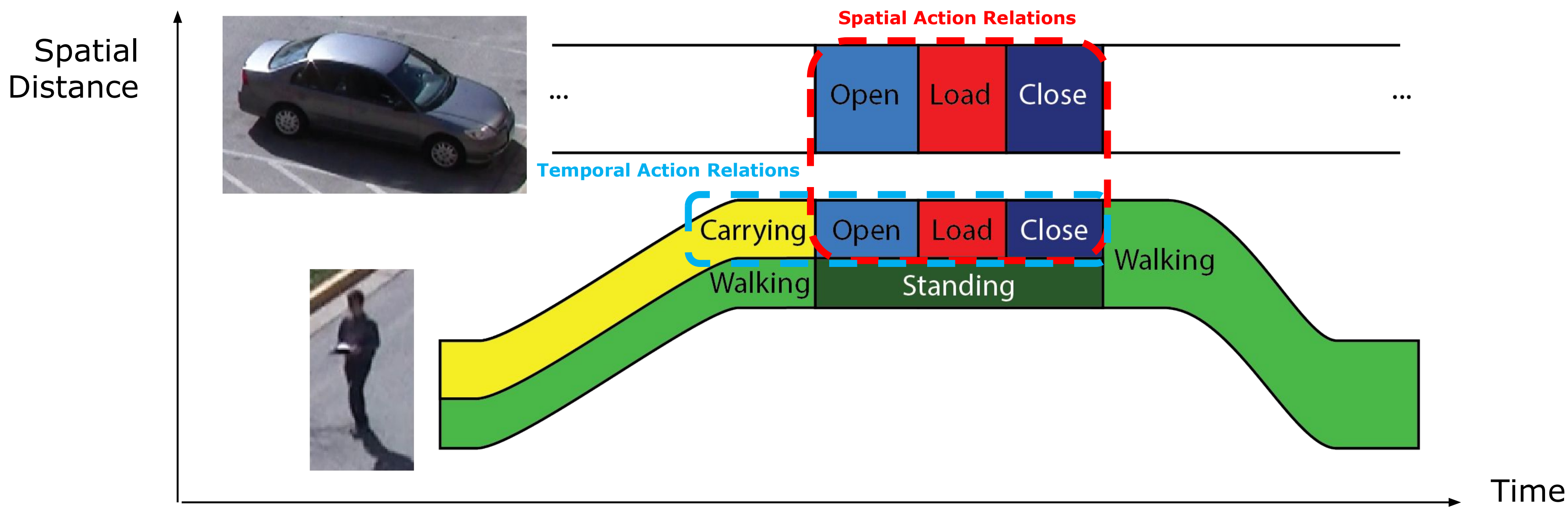
Temporal Action Localization

- Track-wise action localization (R-C3D)
  - Temporal localization and classification of actions
  - Independent of spatial information
  - Generally very dense, many false positives

System Framework

Relation Modeling

- Visually similar actions can be characterized by their spatial proximity to other actors and temporal sequence with other actions
- Modelling as spatial and temporal relations respectively can allow filtering out or correcting erroneous detections



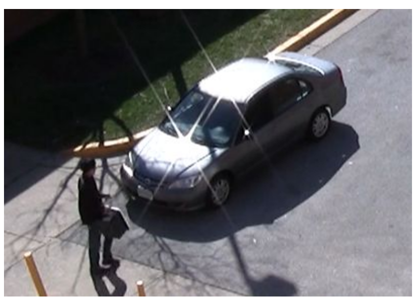
System Framework

Temporal Relation Modeling

- Model the sequences of actions that occur frequently in the dataset
- Heuristic approach:
  - Calculate probability of sequence pairs (X followed by Y) in training set:

$$p_a(X, Y) = \frac{n_{X \rightarrow Y}}{n_X}$$

  - Make set of pairs with probability above certain threshold t_p
  - Penalize proposals with sequences not contained in this set, rescoring them by a factor α, 0 < α < 1

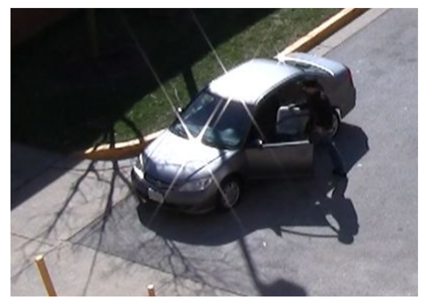


Carrying

$\xrightarrow{p}$ 0.82

Opening

$\xrightarrow{p}$ 0.91



Loading

$\xrightarrow{p}$ 0.27

Entering

System Framework

Spatial Relation Modeling

- Model the actions that expect spatially close objects
- Heuristic approach:
  - List actions that assume actor interactions (person x vehicle, person x person)
  - For all proposals within this list, find the closest proposal of relevant actor class with overlapping boundary boxes
  - Synchronize paired proposals, remove those without valid pairs

Experiments
Overview

- Experiments conducted on each stage of the system
  - Object detection on VIRAT object bounding boxes
  - Action localization on VIRAT actions with ground truth object tracks
  - Relation modeling heuristics on action localization proposals
- Submitted ActEV runs with the 4 most promising combinations
  - Two variants of temporal action localization network, one with a single sampling rate of 6 fps and one with two sampling rates of 6 and 15 fps
  - For each, one submission with basic spatial relation modeling (merging paired proposals) and one with full modeling (also removing proposals with no valid pairs)

Experiments

Results

- Multi sampling rate for temporal action localization leads to some increase in performance, but also lengthens processing time
- Temporal modeling has very slight increase at high thresholds, but fluctuates due to unreliability of probability calculation
- Spatial modeling leads to lower performance, with imprecise time boundary synchronization lowering basic SRM and invalid proposal removal slightly increasing full SRM

	mAP
Single-rate	0.183
Multi-rate	**0.212**

Temporal action localization results

t_p	mAP
1.0	0.213
0.9	0.212
0.8	**0.214**
0.7	0.213

Temporal relation modeling results

	mAP
Basic SRM	0.206
Full SRM	**0.209**

Spatial relation modeling results

Tokyo Tech

Experiments

Results

- Results on official leaderboard
  - Multi-rate sampling (TTA-SF2) improves performance over single-rate (TTA-baseline) as expected
  - Full spatial relation modeling (TTA-SRM) decreases performance to basic SRM (TTA-baseline), contrary to individual results
  - Full SRM (TTA-SF) still leads to lower performance on multi-rate sampling, despite higher amount of false detections removed

	Partial AUDC	Mean p-miss
TTA-SRM	0.85508	0.83174
TTA-SF	0.83456	0.80451
TTA-Baseline	0.81868	0.78228
TTA-SF2	**0.79753**	**0.75502**

Leaderboard submission results

	Basic SRM	Full SRM
Single-rate sampling	TTA-Baseline	TTA-SRM
Multi-rate sampling	TTA-SF2	TTA-SF

Submissions overview

Conclusion

- Experiments on temporal action localization network multi-rate sampling resulted in only notable performance increase
- Experiments on relation modeling didn’t achieve hoped results
- Heuristic approach too naive or too weak to produce significant improvements
- Future works into neural network based temporal relation modeling

Tokyo Tech

Thank You